Expert estimates for feature relevance are imperfect

Patrick M. de Boer
University of Zurich
Binzmühlestrasse 14
8050 Zurich
pdeboer@ifi.uzh.ch

Marcel C. Bühler
University of Zurich
Binzmühlestrasse 14
8050 Zurich
marcel.buehler@uzh.ch

Abraham Bernstein
University of Zurich
Binzmühlestrasse 14
8050 Zurich
bernstein@ifi.uzh.ch

***Abstract*—An early step in the knowledge discovery process is deciding on what data to look at when trying to predict a given target variable. Most of KDD so far is focused on the workflow after data has been obtained, or settings where data is readily available and easily integrable for model induction. However, in practice, this is rarely the case, and many times data requires cleaning and transformation before it can be used for feature selection and knowledge discovery. In such environments, it would be costly to obtain and integrate data that is not relevant to the predicted target variable. To reduce the risk of such scenarios in practice, we often rely on experts to estimate the value of potential data based on its meta information (e.g. its description). However, as we will find in this paper, experts perform abysmally at this task. We therefore developed a methodology, *KrowDD*, to help humans estimate how relevant a dataset might be based on such meta data. We evaluate *KrowDD* on 3 real-world problems and compare its relevancy estimates with data scientists' and domain experts'. Our findings indicate large possible cost savings when using our tool in bias-free environments, which may pave the way for lowering the cost of classifier design in practice.**

***Index Terms*—None**

I. Introduction

A critical step in knowledge discovery takes place in the very beginning: the selection of the data to be used. In practice, the final dataset used is an amalgamation of many different attributes, scattered across company-internal data and public data. For example, in order to predict the sales of a pharmacy, one might want to acquire historical sales data for each day, and try to predict these using features like the number of employees working at a given day (internal data), the outdoor temperature (external), the current severity of the flu (external), etc. Acquiring the right data is critical for data science projects, since the performance of the actual feature selection and model induction is dependant on the human decision what datasets to acquire in the first place. In practise, this decision often takes place in workshops, where project members decide on what internal and external data sources to tap for feature selection. Since data sources vary in their cleanliness, structure and price, each comes at different cost, which entails a cost-payoff calculation for the selection of data sources to acquire. Therefore, an implicit outcome of such workshops is a (human) ranking of the expected relevancy for the potential features engineerable from the available data sources.

However, as we find in this paper (see Section II), human relevancy estimation of individual features is often unreliable. This constitutes an important problem for data science teams: For topics where literature or experience listing relevant data is unavailable, data science teams are essentially in the dark as to what data sources to acquire.

In Section III, we therefore propose an early method to more accurately estimate feature relevancy than the current industry standard — without any more information than the names and semantic descriptions of the features it contains. Our method, called *KrowDD*, is based on crowdsourcing, where we aggregate many people's estimates to approximate a (potential) feature's Information Gain, before the actual data for that feature is available.

When compared to domain experts and data scientists, our method performs significantly better at estimating feature value in case of a non-skewed target variable. Our contributions in this paper therefore are:

1) a study quantifying the performance of human feature relevancy estimation in practice
2) a method to estimate feature relevancy only using meta-information (name and description)

II. Quantifying Expert Performance

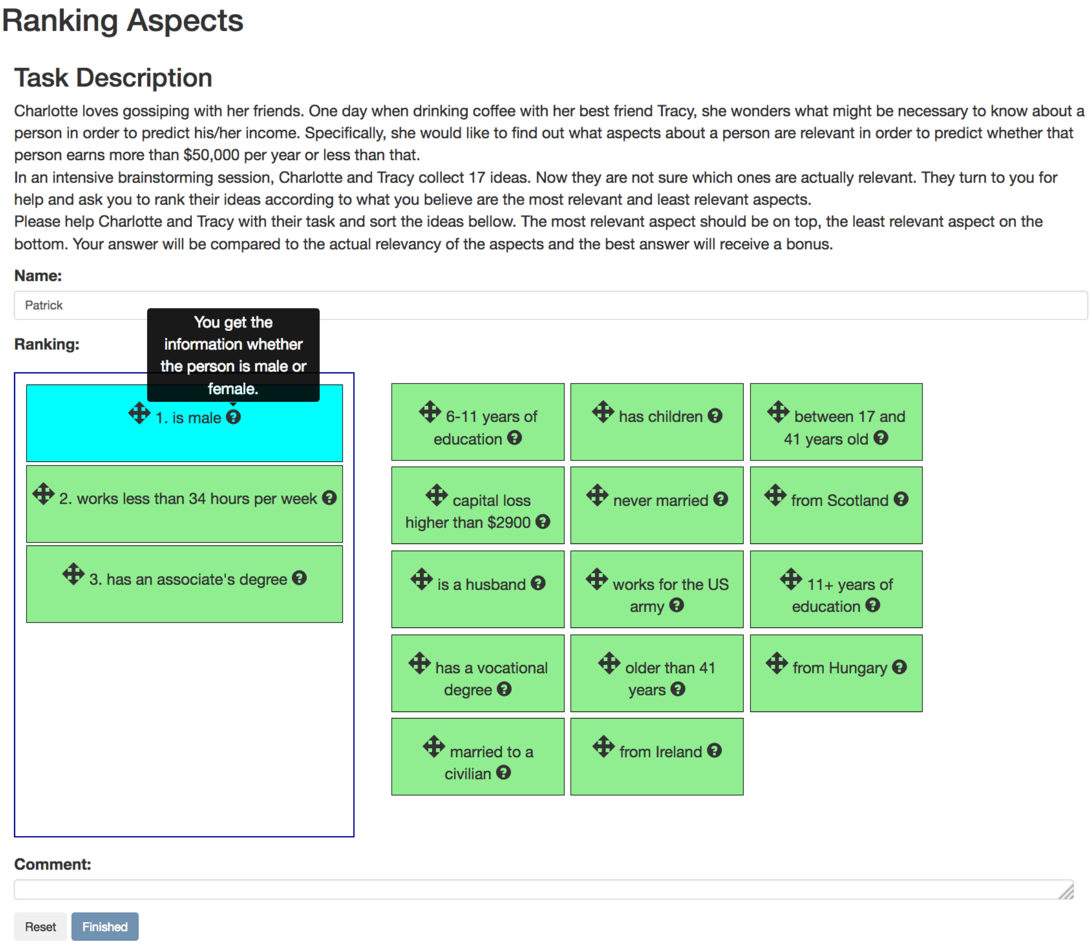
Imagine, you'd need to build a classifier to predict the sales of a pharmacy for 7 days into the future. The pharmacy supplied you with some internal data for past sales, including features such as the number of employees working on a given day, the number of products on display in a given branch, etc. Your initial model doesn't yield stellar performance, and you decide to try improving it by integrating public data. You come up with a few ideas, such as data on the current spread of the flu virus, data for school holidays as well as weather data. Obtaining all of these data requires time, which is scarce. You therefore rank your ideas based on how relevant you deem them to your target variable (sales) and then start acquiring, cleaning and integrating these datasets.

In practice, the process of ranking such potential data sources is either conducted by data scientists directly, or in consultancy with a domain expert (”subject matter expert”). In case there is no publicly available report on what data worked well to predict a target variable, data scientists need to rely on their intuition to determine where to invest their time first. Given that the obtained data in this very first phase ultimately limits the performance of any model induced on it, it would be important to quantify how well human intuition — the de-facto standard — works with estimating data relevance.

In existing literature, we could only find anecdotal evidence [1] quantifying the performance of the current state of the art as rather suboptimal. A more fine-grained view would be to look at individual features of such external datasets; as datasets might primarily be obtained for one or more of the features they contain.

We therefore set out to answer *how well can people rank features according their relevance to a target variable*.

The question was structured after the sequential model data scientists employ to acquire more data, where the most promising feature (i.e. the one with highest expected relevance) is obtained first, followed by the second, third, etc. In order to most accurately represent the real world, we included both prevalent kinds of human experts employed in practice: data scientists and domain experts. Test subjects of each group were asked to rank features according to their relevance for the target variable. These two groups were then compared to a third (synthetic) group of test subjects, where each synthetic "test subject" would return a randomly ordered list of the supplied features. This third group will be called *random*. To compare and contrast these 3 groups, each test subject's ranking needs to be quantified in terms of its performance. A common way to evaluate the performance models in literature is by comparing their Area under Curve when using Receiver Operating Characteristics (ROC AUC) first presented by Provost et al. [2]. In order to compare the feature rankings of individuals, we therefore induce a classifier for each feature count for a test subject's ranking. For example, if one person ranked the available features A, B, C as *"A,C,B"*, this would entail training 3 classifiers: the first classifier only using feature A, the second using features A and C and the third using features A, C and B. Using these measures, all groups can be compared at each feature-count (e.g. for two: a comparison of all groups when only using the two features ranked highest). This method of comparing feature rankings is similar to the one proposed by Slavkov et al. [3].



Ranking Aspects

Task Description

Charlotte loves gossiping with her friends. One day when drinking coffee with her best friend Tracy, she wonders what might be necessary to know about a person in order to predict his/her income. Specifically, she would like to find out what aspects about a person are relevant in order to predict whether that person earns more than $50,000 per year or less than that.
In an intensive brainstorming session, Charlotte and Tracy collect 17 ideas. Now they are not sure which ones are actually relevant. They turn to you for help and ask you to rank their ideas according to what you believe are the most relevant and least relevant aspects.
Please help Charlotte and Tracy with their task and sort the ideas bellow. The most relevant aspect should be on top, the least relevant aspect on the bottom. Your answer will be compared to the actual relevancy of the aspects and the best answer will receive a bonus.

Name: Patrick

Ranking:

You get the information whether the person is male or female.

Comment:

Fig. 1: Drag & drop interface to elicit a test subject's feature relevancy estimations for a given topic. This figure displays the *Income* condition, where a person has already ranked three features. Test subjects would drag a feature (green box) from the right part of the screen onto the left part of the screen at the position perceived as corresponding to the feature's rank. The screenshot shows the simplified vocabulary used to elicit estimates of people without background in Data Science.

A. Experiment Setup

An important realisation based on our method of comparison is its model-dependency: the AUC calculated on the 2 highest ranked features with a Naive Bayes classifier might differ from the AUC of the same features with a Decision Tree. We mitigate this problem by choosing a classifier that is very common (Naive Bayes) and then investigate whether other popular classifiers would have lead to different results. Note, that in the field of feature selection, which is closely related to our problem, inducing classifiers with a given selection of features for its comparison is quite common (e.g. [4]).

In order to support our results, we conduct our experiments on three different datasets (”conditions”) to generalize. We chose datasets that are public domain, well known in the ML community and for which no highly specialized knowledge would be required. An example for specialized knowledge to be avoided could be, predicting the likelihood of breast cancer based on presence of BRCA 1 or 2 gene test outcomes.

In preliminary experiments using simulations, we found our method to be sensitive to skewed target variables. We therefore chose 2 datasets with balanced target variable distributions (*Student* and *Olympics*) and one dataset with a skewed target variable (*income*). For each dataset, we included roughly as much relevant features as irrelevant ones, whereas we included the N highest-ranked features (by their Information Gain) and the N lowest-ranked features. Table I shows the datasets and the number of relevant features each.

To simplify our task descriptions for humans, we preprocessed each dataset by binning all numeric variables into 3 equally sized bins (low, medium, high). Rows with more than 50% with missing values in our target features were dropped, the remaining missing values were imputed by their means. Simulating a real-world experience, we created comprehensive (semantic) descriptions of each feature and included them with their names as meta information. For example, the description for the third bin (*high*) of the electricity consumption feature used to predict whether a country receives at least one medal in the Olympics: *You get the information whether the country has a high electricity consumption per person (more than 4200kWh per person) (YES) or less than that (NO).* Using these descriptions, we created a website, where people could submit their ranking for feature relevancy for a given condition, by putting the most relevant feature on top, followed by the second, third, etc. through drag-and-drop. The descriptions of each feature could be accessed by hovering over the corresponding question mark icon. Figure 1 shows the user interface presented to our test subjects for this task. The test subjects would drag features from their unranked set on the right-hand side of the screen to their ranking on the left-hand side. The unranked set on the right is shuffled for every participant to avoid biasing the outcome.

The Area under Curve (AUC) was calculated as the average of a 10-fold cross validation for each test subject's ranking and each cumulative feature count. We then compared the confidence intervals of each group. Confidence intervals were calculated through Efron's bias-corrected and accelerated bootstrap [5].

For each dataset (”condition”) we used 10 experienced data scientists and between 10 and 18 domain experts depending on the dataset and expert availability.

Data scientists were recruited through the freelancer platform Upwork[1], at the time of writing one of the largest pools of online workers for data science type tasks. We limited our search to data scientists who reported experience in Machine Learning, Data Science or Data Analytics through their respective tags or descriptions. Data scientists were paid $10 for participation, which at the time of writing is a common pay for small-scale data science jobs like feature selection. They were then asked to rank variables for all conditions, whereas the learning effect was mitigated by randomly permuting the order of the conditions for a given worker. All data scientists were asked to additionally answer a questionnaire, in order to report on their previous experience and education. Domain experts were Swiss and German volunteers, selected based on their experience and profession.

1) Condition 1: Olympics: In the *Olympics* dataset, our (binary) target variable is, whether a country received at least one medal in the Olympics or not. To build a model, we acquired yearly-data for each country that participated in the Olympics since 1996. Following Bredtmann et al. [6]'s paper, we included ”region” as one of the relevant features. Additionally, we acquired the following features from the *CIA World Factbook* for each of these countries where available: education expenditures, inflation rate, unemployment rate, public debt, electricity consumption, exports and share of internet users. From these features, we selected 8 binned features with an Information Gain higher than 0.08 into the *relevant* group and 5 features with an Information Gain lower than 0.02 into the *irrelevant* group.

Out of 46 approached domain experts, 10 submitted their answers (response rate 22%). Among them 7 athletes (1 of which actually won gold in the Paralympics) and three representatives of the sports departments in a city/state or national government.

2) Condition 2: Student: Our second condition focused on the (binary) target variable of whether a Portuguese high school student would finish a Portuguese language class with a grade above the classes' median or below. The dataset as well as all of its features were acquired from the UCI Machine Learning repository [7].

For this condition, we chose our features from the variables *Fjob* (father's job), *Medu* (mother's education level), *Mjob* (mother's job), *failures* (number of classes failed), *paid* (does the student take paid extra classes?), *studytime* (number of hours spent studying besides school, *famsize* (family size), *health* (health status), *Pstatus* (parent relationship status), *absences* and *age*. In total, we chose 8 binary features with an Information Gain higher than 0.01 and 7 features with an Information Gain lower than 0.002.

We were able to recruit a total of 18 Swiss or German high school language teachers as domain experts for this condition. 14 of them were directly approached by us, out of which at least two have forwarded our emails to their colleagues.

3) Condition 3: Income: In our third condition, we sought a skewed (binary) target variable, where significantly more than half instances of the dataset would share the same value of the target variable. A well-known dataset with this property is 'Census Income', also available on the UCI Machine Learning Repository [7]. In most societies — and in particularly in the US — income is a skewed variable, where a few wealthy individuals dramatically skew the average income figure of a

[1] https://www.upwork.com

TABLE I: Datasets used for the comparison between *human experts* and *random*. For each dataset we included roughly the same number of relevant and irrelevant features (as judged by their Information Gain)

	Olympics	Student	Income
Size (rows)	839	649	32561
# Relevant features	8	8	10
# Irrelevant features	5	7	7

Note: These numbers represent the data set after cleansing and binarization.

country. The main reason to include a skewed target variable, was to explore the influence of the base rate problem described by Kahneman [8] on expert's judgments.

In this condition, we chose features describing *marital.status*, *relationship*, *education.num* (number of years spent for education), *sex, age, hours per week* (hours spent at work), *native.country, education, occupation* and *capital loss*. We used 17 binned variants of these features, among them 10 with an Information Gain higher than 0.02 and 7 with an Information Gain lower than 0.00008.

We were able to recruit a total of 12 domain experts for this condition (response rate 60%). Among them were 2 tax experts, 6 professional fiduciaries and 4 leaders of companies or department leaders of big companies.

B. Demography

The 20 data scientists participating in this experiment went through university education and possess at least a Bachelor of Science (10 Master of Science, 1 PhD). Two data scientists were female, 18 Male. 10 were from Asian countries, 5 from Europe, 2 from North America, 2 from Africa and 1 from South America. They were between 20 and 39 years old with an average of 29 years. Experience varied between 1 year and 13 years with an average of 4.5 years of experience. The Kendall-Tau correlation between experience (in days) and ranking performance (in terms of AUC) was 0.019 with P=0.83 for absence of an association. The low correlation suggests, that data scientists with more experience did not perform better in ranking feature's by their relevance than data scientists with less experience.

C. Results

Figure 2 shows the results of the experiment in all conditions. Strict superiority, the case where the confidence interval of one group is clearly higher than the one of another group, is not visible. In fact, all groups seem to perform remarkably similar.

Table II shows our comparison of the groups using Welch's t-test for unequal variance [9] and calculated effect sizes using Hedges' g [10] for small sample sizes.

Figure 3 sheds light on the relative performance of experts across all three conditions. In this figure, we used linear interpolation to normalize the AUC for each group to the range [0, 1], whereas 1 is the highest AUC and 0 the lowest AUC achievable for a given number of features.

TABLE II: Comparing *human experts* with *random*: Hedges' g effect size [10] for 6 features

	Data Scientists	Random
Domain experts	0.283^{+}	0.508*
Data scientists		0.296^{+}

(a) Student dataset

	Data Scientists	Random
Domain experts	0.150^{+}	-0.045^{+}
Data scientists		-0.207^{+}

(b) Income dataset

	Data Scientists	Random
Domain experts	-0.769^{+}	-2.002**
Data scientists		-0.953**

(c) Olympics dataset

Note: + indicates $P > 0.05$, * indicates $P <= 0.05$ and ** indicates $P <= 0.01$. P-values were calculated using Welch's t-test for unequal variance [9].

No group significantly outperforms another in the relative case either. We therefore conclude, that we can not observe a significant improvement of using any group of experts over another in our targeted conditions (subject to the limitations outlined below). In all conditions with balanced target variable distributions, we find *random* to not significantly differ from experts. In case of *Income*, there is a significant difference judging by the t-test, albeit with low effect size (Hedges' g).

D. Discussion

In the results above, we observe that experts (data scientists and domain experts) ranking of features according to their relevance is not much better than random selection. This finding needs to be related to literature in order to better understand the reason for this effect.
We conjecture, that human biases might be responsible for the abysmal result of experts. It has been established by various well-known psychologists, that expert judgments are subject to human biases, may be poorly calibrated or could be self-serving (e.g. Tversky and Kahnemann [11], Krinitzsky [12]). This leads to systematic bias being present in expert judgments.

A prominent example for such a bias leading to misjudgments is the *availability bias* originally introduced by Tversky and Kahneman [13]. It is used to describe the phenomenon of overestimating the probability of an outcome based on the ease with which it comes to mind. If domain experts or data scientists have not gathered prior experience on the relevancy of a feature to a target variable, they might defer to the availability heuristic to estimate the likelihood of a feature being influential or not. The availability bias can extended by the *overconfidence* bias, i.e. that the confidence people have

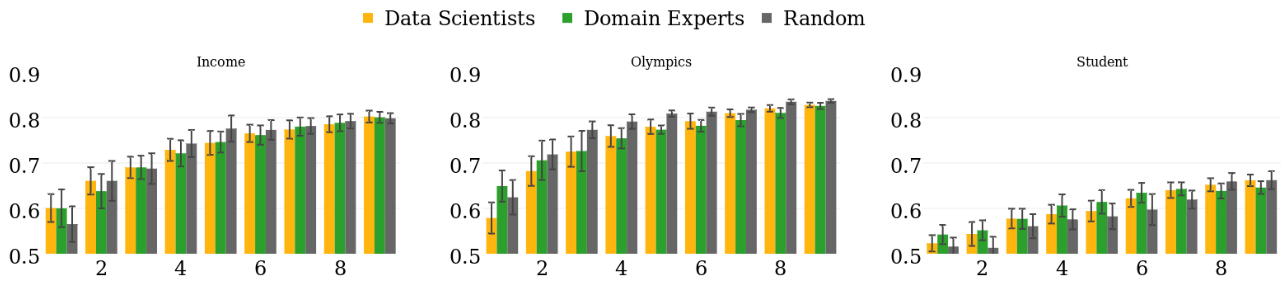


Fig. 2: *Human experts* compared against *random*. The x-axis denotes the number of features involved in training the classifier, while the y-axis show the range of AUC (including a 95% confidence interval) produced when training with said number of features. There is no strict superiority by any group visible, which suggests that experts are not significantly better than *random* at ranking feature according to their relevance.

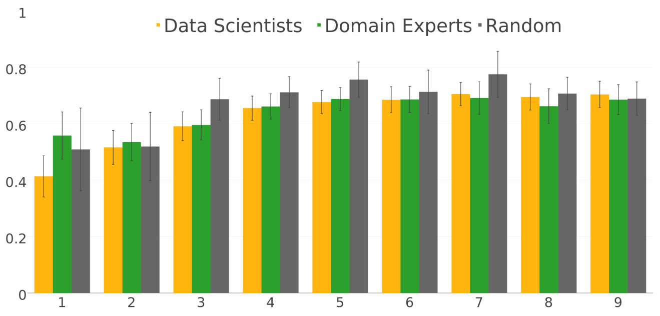


Fig. 3: Performance on test subjects across all conditions combined. y-axis is the normalized AUC of each condition (higher is better), with error bars for 95% CI, whereas the x-axis denotes the number of features used for training. This figure shows the average human performance relative to the respective maximum / minimum of each dataset. No clear difference is visible in the relative case either.

in their judgment tends to be higher than its accuracy — a phenomenon that becomes stronger with more difficult tasks.

Given the sheer size and the complexity of the existing literature on feature selection, one could safely conclude that the task of ranking features according to their relevance could be considered rather difficult for humans. The two biases outlined above are well accepted and affect experts in many fields with human decision making.

For example, in risk analysis, experts are polled to estimate the confidence interval of a given event to occur. McKenzie et al. [14] find, that people's 90% confidence interval contain a true value about 50% of the time. They observed similar levels of overconfidence in experts and laymen (46% vs 49%). In finance, Oberlechner et al. [15] find expert currency dealers to be overconfident, despite the evolutionary argument that irrational traders would sustain continuous losses and be pushed out of market eventually and despite the immediate and accurate feedback for judgments cast. In medicine, Blendon et al. [16] surveyed 831 physicians and 1207 patients on the extent of a possible medical errors that they or their immediate family members experienced. Physicians reported 35% of such errors, while patients reported 42%, indicating possible overconfidence on the side of physicians (or overt sensitivity of patients). A similar gap is visible when investigating the health consequences of such medical errors, which physicians report to be serious in 18% of cases, whereas patients see this figure at 24%. Berner and Graber [17] review medical papers on diagnostic error due to overconfidence in different areas of medicine, ranging from fields depending on visual interpretation (radiology and pathology), where errors seem to be lowest (ranging from 2%-5%), to the clinical specialities, where diagnostic error is approximated to be 15%. The presence of these errors seems to be stationary over time, which might be attributed to Tetlock's [18] finding, that experts are less likely to change their minds than non-expert, when new evidence disproving their beliefs appears.

The argued-for difficulty of our task at hand, paired with known human biases — such as overconfidence — particularly applying to experts, might have lead to the performance of experts not surpassing random.

E. Related Work

Our method is based on literature in the fields of Feature Selection and Crowdsourcing.

Feature Selection is used to (i) avoid overfitting through dimensionality reduction, (ii) providing faster and more cost efficient models and (iii) to gain deeper insight into the underlying processes generating the data [19]. Kohavi et al. [20] proposed to categorize existing techniques into wrappers and filters, whereas wrappers learn a classifier in tandem with feature selection. By contrast, filters are classifier-independent and are typically used before classifier induction. Popular examples for filters include Information Gain [21], the Pearson Correlation and Markov blanket filters [4]. Information Gain belongs to the univariate filters, and is based on variable entropy [22]. Averaging independent estimates of many different people, a hypothesis termed Wisdom of Crowds, has been shown to be relatively accurate in some conditions [23].

To the best of our knowledge, existing feature selection techniques require data for the features under processing to be present to calculate metrics on it.

Our method, by contrast, does not require data to be present at the time of processing. Instead, our method is based on estimating the conditional entropy of the features part of feature selection.

For example, Francis Galton analyzed the accuracy of individual estimates of an ox's weight at a regional fair [24]. He found, that the average estimate (1197 pounds) was off by just 1 pound (the actual weight of the ox was 1198 pounds). Treynor [25] ran two bean jar contests, where he let 46 and 56 students estimate the number of beans in a jar. In the first experiment, the estimate diverged by roughly 4% (810 actual beans, 841 estimated), while in the second experiment it diverged by roughly 2% (850 actual beans, 871 estimated). James Surowiecki [23] established the conditions for the Wisdom of Crowds to work when (i) knowledge about the cause is available, (ii) the crowd is motivated to be accurate, (iii) the crowd is independent and (iv) diverse. Simmons et al. [26] provide an excellent overview over the Wisdom of Crowds hypothesis and its conditions. Additionally, they look at the impact of systemic biases present in individual decision making.

In the context of crowdsourced feature selection, Cheng et al. proposed Flock [1], a method and system for crowdsourced feature generation. Flock is a platform to guide crowds to nominate features and provide labels for them. To nominate features, Flock lets crowd workers compare positive and negative examples of the binary target class. Crowd workers are then asked to state a reason how they differ. These reasons are later clustered into features, followed by recruiting crowd workers to provide labels for the clustered features for the remainder of the dataset. Another approach based on comparing examples was proposed by Zou et al. [27], where crowd workers are asked to name a feature common to two out of three displayed examples, followed by providing labels for the nominated feature. Both approaches are based on the crowd providing labels for the nominated features in order to induce a classifier. This requires asking crowd workers to label the full dataset for the nominated features — a potentially costly endeavor that does not scale well to larger datasets. Besmira et al. [28] propose to use budgeted learning when labelling the training set and the test set for feature selection. Our approach avoids this problem altogether, as it does not depend on crowd labelling.

III. Possible Remedy: KrowDD

In this section, we will present *KrowDD*: a method to estimate feature relevancy to a target variable only by knowing meta-data (name and description) of the feature, i.e. without access to the feature's data. Such a method can be helpful when estimating the usefulness of datasets for Knowledge Discovery applications; before running an actual feature selection algorithm.

The idea is based on approximating values used to calculate the Information Gain, a common feature selection method. More specifically, we let the crowd estimate the conditional means used in the calculation of the conditional entropy of a variable. When applied to a set of features, an approximation of an order by relevancy arises. Since the complexity of Information Gain grows linearly with the number of features part of the analysis, our method scales linearly to large numbers of potential external datasets.

Information Gain measures the information obtained for predicting a target variable by knowing the value of a feature variable. The Information Gain for a target variable Y given a feature X can be calculated by

$$IG(Y|X) = H(Y) - H(Y|X)$$

Whereas $H(X)$ denotes the entropy [22] of a variable X calculated by

$$H(X) = -\sum_i p(x_i) \cdot \log_2 p(x_i)$$

and $H(Y|X)$ denotes the conditional entropy of Y given X calculated by

$$H(Y|X) = -\sum_j p(x_j) \cdot \sum_i p(y_i|x_j) \cdot \log_2 p(y_i|x_j)$$

The conditional entropy of $H(Y|X)$ denotes the expected number of bits needed to transmit a variable Y if the other party knows the value of X.

For the case of binary variables, the term to calculate the Information Gain can be simplified, such that it only needs values for $P(Y=1|X=0)$, $P(Y=1|X=1)$, $P(Y=1)$ and $P(X=1)$. Their counter parts can be calculated by subtracting the variable from 1. Note, that categorical variables can be transformed to binary variables through dummy extraction. Numerical variables can be transformed to categorical variables through binning at a user-defined loss of precision.
Our method is based on estimating the value for these variables through the median of a number of crowd estimates. As shown in the related work section, averaging many crowd estimates is commonly used in literature to estimate unknown parameters. Applied to our case, an estimation of the share of entries in a variable is required, where a certain property is true: $P(Y=1|X=0)$. Practically, one could ask a crowd of people questions in the following format: *"What's the share of Y having $X=0$*. For example: "What's the share of countries winning at least one gold medal in the Olympics, which consume less energy per capita than average)."
Crowd estimates through averaging are inherently noisy and imperfect. It is therefore necessary to quantify the error associated with noise in our validation.

IV. Validation

We evaluated our method by reusing the real-world datasets (conditions) introduced in Section II. Based on our finding, that the different human experts surveyed in Section II did not diverge significantly from each other, we combined their judgments. This allows us to compare *KrowDD*'s performance directly with human experts and reduced cognitive load.

A. Experiment Setup

In order to approximate Information Gain following the *KrowDD* approach, we elicited crowd estimates for the following meta-data for each feature:

- $P(Y = 1|X = 0)$: The probability of the target given the feature was False.
- $P(Y = 1|X = 1)$: The probability of the target given the feature was True.
- $P(X = 1)$: The (prior) probability of the feature variable being True.

For each feature, we acquired at least 16 estimates for this meta-data priced at \$0.10. The sample size of estimates per feature is stated as lower threshold (of 16 estimates), since repeated answers of the same workers were removed post-collection. Estimates for $P(Y = 1)$ (the prior of the target variable) were priced at \$0.04 and sampled using the same strategy. All estimates were obtained from crowdworkes recruited through Amazon Mechanical Turk (AMT)[2], where we limited our selection to experienced[3] US workers. AMT samples were collected on workday mornings Eastern Time in parallel for a given condition. $P(Y = 1)$ was already known (as 0.5) in conditions, where we picked the target variable to be symmetrically distributed (*Olympia* and *Student*). Only in *Income*, we turned to the crowd to find it.

To compare *KrowDD* with our human judgments obtained in Section II using their confidence intervals, we subsampled 9 crowd answers for the variables required to calculate *KrowDD*'s Information Gain (without replacement). Subsampling allowed us to generate multiple estimates for the AUC created by *KrowDD*-guided feature selection, such that we could calculate its confidence interval. More specifically, we iteratively subsampled 9 crowd estimates with replacement for each feature and conditional mean. We repeated the process above 19 times, resulting in 19 AUC scores per feature per condition for *KrowDD*. The confidence intervals were calculated by the bias-corrected and accelerated bootstrap introduced by Efron [5] using 10,000 bootstrap samples.

B. Results

Figure 4 compares *KrowDD*'s performance with human experts for a given number of features (x-axis). One can see *KrowDD* to yield better results in cases with balanced target variable (*Olympics* and *Student*), but performing slightly worse in case of skewed target variable (*Income*). In *Income*'s case, *KrowDD*'s 95% confidence interval overlaps with human expert's, whereas its mean performance is lower. In cases with balanced target variable, *KrowDD*'s confidence interval does not overlap with expert's for most feature counts (x-axis), passing a simple CI overlap-test, which suggests CI-superiority of *KrowDD*. The difference between human experts and *KrowDD* in all cases is quantified and documented in table III. Figure 4 also shows the reference performance of the best possible feature selection and worst possible selection for each number of features, which were established through exhaustive search.

TABLE III: Comparison between *KrowDD* and *human experts*: Hedges' g effect size [10] for 1-9 features. Cases where *KrowDD* performs better than humans are highlighted in bold. *Student* and *Olympics* saw tend to show better performance of *KrowDD* than *human experts*, *income* shows *KrowDD* to perform worse.

	Student	Income	Olympics
1	**0.707****	-0.959**	**1.606*****
2	**0.788****	-0.443^{+}	**1.448*****
3	**0.719***	-0.285^{+}	**1.375*****
4	**0.773****	-0.499^{+}	**1.716*****
5	**0.899*****	-0.737**	**2.167*****
6	**0.435**$^{+}$	-0.711*	**1.864*****
7	**0.560**$^{+}$	-0.780**	**2.105*****
8	**0.660***	-0.422^{+}	**1.812*****
9	**0.415**$^{+}$	-0.133^{+}	**1.901*****

Note: + indicates $P > 0.05$, * $P <= 0.05$, ** $P <= 0.01$ and *** $P <= 0.001$. P-values were calculated using Welch's t-test for unequal variance [9].

In order to estimate a feature's Information Gain, *KrowDD* samples estimates of crowd workers for the variables described above. We calculated our AUC, by using 9 crowd estimates per variable. While we have selected the number 9 experimentally before running the actual evaluation, some guidance as to the number of samples per variable necessary would be helpful for applying *KrowDD* in practice. Figure 5 therefore compares the absolute difference between the aggregate estimate of a variable and its actual value across all conditions. For example, if 5 crowd workers estimated $P(X_4 = 1|Y_1 = 0) = 0.64$ (by their median estimates for $P(X_4 = 1|Y_1 = 0)$) and the actual value of $P(X_4 = 1|Y_1 = 0) = 0.7$, the delta would be 0.06. Note, that $P(X_4 = 1|Y_1 = 0)$ designates the conditional mean of the fourth feature in the first condition. The figure shows, the accuracy of the aggregated crowd judgment improves as the number of judgments increases until roughly a sample size of 6 is reached, suggesting that 6 crowd-samples per variable might be enough.

V. Discussion

Given the results of *KrowDD*'s performance, we see first evidence for two findings: *KrowDD* is vulnerable to skewed target distributions; and in case of balanced target variable distribution, it outperforms human experts. Particularly the case of skewed target distributions is interesting, as it reproduces a finding of Kahnemann and Tversky's landmark paper on the psychology of prediction [29]. In their paper, Kahnemann and Tversky find that people predict outcomes by their representativeness of evidence — prior probability of the outcome is systematically ignored. Applied to *KrowDD*'s case, crowd workers asked to estimate $P(Y = 1|X = 0)$ and $P(Y = 1|X = 1)$ could not take the prior probability of Y

[2] http://mturk.com

[3] Workers with more than 4000 approved answers (called HIT) and a total of less than 4% rejections

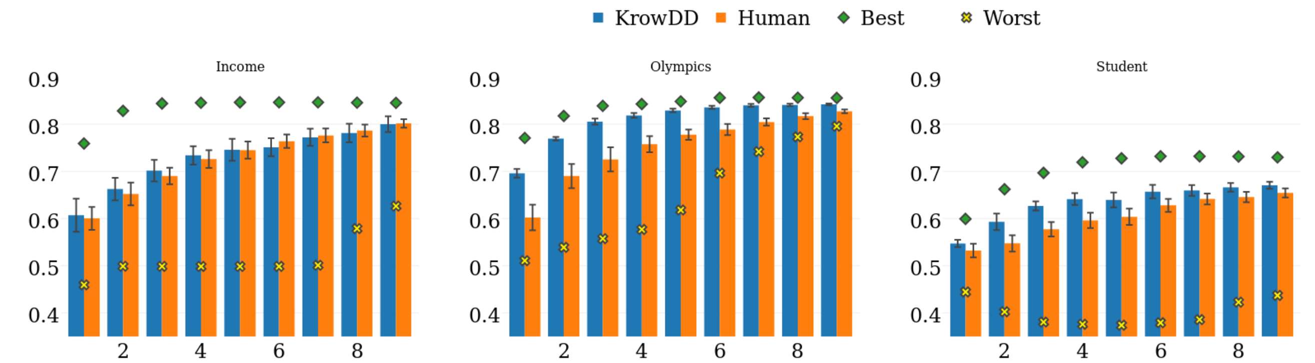


Fig. 4: Average AUC and confidence intervals for *KrowDD* and *human experts* (*data scientists* and *domain experts*). The x-axis denotes the number of features used to train the classifier and the y-axis the average AUC with a 95% confidence interval (higher is better). The graph suggests *KrowDD* to perform better than *human experts* in cases with balanced target variable, and worse in case of a skewed target variable.

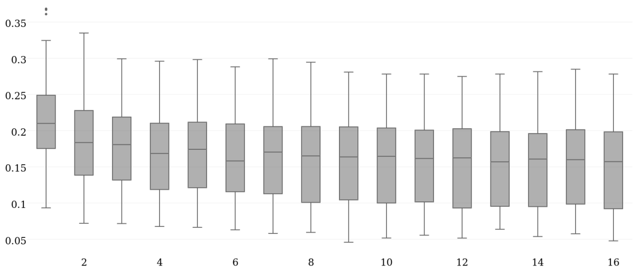


Fig. 5: Combined crowd errors for $P(X=1)$, $P(Y=1|X=0)$ and $P(Y=1|X=1)$. For each data point, we sampled 19 times without replacement. The x-axis denotes the number of answers samples in each iteration and the y-axis the delta to the actual values.

TABLE IV: Comparing *KrowDD* with the ranking performance of crowdworkers from AMT: Hedges' g effect size [10] for 1-9 features. Cases where *KrowDD* performs better than laypeople are highlighted in bold. *Student* and *Olympics* again show *KrowDD* to be superior, while human intuition seems to work better in *Income*.

	Student	Income	Olympics
1	**1.439*****	-1.029**	**2.156*****
2	**0.918****	-0.547^{+}	**2.294*****
3	**0.919****	-0.164^{+}	**1.867*****
4	**1.334*****	-0.270^{+}	**1.721*****
5	**1.256*****	-0.999**	**2.077*****
6	**0.616**$^{+}$	-0.705*	**2.281*****
7	**0.660***	-0.737*	**1.786*****
8	**0.879****	-0.326^{+}	**1.573*****
9	**0.884****	**0.195**$^{+}$	**1.966*****

Note: + indicates $P > 0.05$, * $P <= 0.05$, ** $P <= 0.01$ and *** $P <= 0.001$.

into account, which might have lead these estimates astray. In the data, we indeed see a significant difference ($P < 0.001$ using a t-test, with large effect size) between the accuracies of estimates of conditional probabilities in the case of balanced vs skewed target variable. The effect sizes (Hedges' g) for estimates for $P(Y=1|X=0)$ and $P(Y=1|X=1)$ when comparing the skewed with the balanced conditions were 5.091 and 4.567 respectively.

Herzog et al. [30] found the crowd to predict outcomes of sport events (soccer and tennis) accurately. This might mean, that crowd workers might inherently perform better at predicting sports events (such as the Olympics) than experts, which could be a possible explanation why *KrowDD* (powered by such crowd workers) performs better than experts. If crowd workers had a systematic knowledge advantage over experts in the investigated conditions, our observed superiority of *KrowDD* in balanced target variable cases would be rendered moot. To shed light on a possible knowledge advantage of the crowd employed by *KrowDD* from Amazon Mechanical Turk, we therefore compared their performance in feature relevance ranking to *KrowDD*'s performance. In essence, crowd workers went through the same ranking procedure as experts have in Section II. Table IV shows the comparison of crowd workers using *KrowDD* and employing lay crowd workers in ranking features directly. We find the same pattern as when comparing *KrowDD* to experts: it performs better in cases of balanced target variables, and worse in case of a skewed one. Based on this finding we are assured, that it was indeed *KrowDD* giving crowd workers from Amazon Mechanical Turk an edge over experts.

Using these findings, we are convinced that *KrowDD* poses a first step on a path improving feature relevancy estimation through meta-data. For the case of a balanced target variable, *KrowDD* might be able to save data science teams a substantial amount of time and resources.

VI. Limitations and Future Work

The method proposed in this paper faces a few threads to its validity.

The experiments in this article focussed on using lay people as drivers for *KrowDD*, due to lower cost and higher availability. More generally, drawing upon the finding of suboptimal domain expert performance from Section II, *KrowDD* might constitute a method with which domain expert knowledge might be put to use in a better fashion than current state-of-the-art. For example, to estimate feature relevancy when predicting the likelihood of breast cancer, a user of *KrowDD* might draw upon physicians' insight as a driver for *KrowDD*.

Another important point to consider is that *KrowDD* is based on Information Gain as introduced by [21]. It therefore inherits Information Gain's advantages and disadvantages. Specifically, ex-ante, *KrowDD* can not estimate the relevance of interactions between two features due to it being a univariate feature selection metric. Closely related, *KrowDD* does not consider feature autocorrelation in its current form: looking for semantically similar variables would lead to similar utility scores (approximated IG), despite the fact that both of them might explain the same variance of the target variable. While it has been shown that autocorrelated features may not be redundant to a classifier [31], *KrowDD* does not discount its relevancy metric for the arising interaction effect.

Another threat to *KrowDD*'s performance might be posed by regional differences. For example, estimating the relevancy of features used to predict whether a German bank would give credit or not might be better answered by a German crowd than a US crowd due to the cultural differences of credit handling.

Lastly, in this article, we only supplied AUC numbers based on a Naive Bayes classifier (which was selected based on its popularity and representativeness). We reran the full evaluation using two other popular classifiers (C4.5 and a Multilayer Perceptron), and found them to not change our main takeaways.

VII. Conclusion

This article outlines two main contributions: it shows that human experts (data scientists, domain experts) do not perform significantly better in selecting relevant datasets than random. In a situation, where a data scientist is tasked with adding external data to improve a model, this finding implicates, that the data scientist's intuition on what data to look for may be flawed. Given systematic flaws in an people's judgments, we conjecture, that data scientists often spend time and resources obtaining, cleaning and integrating data for ineffective features.

Our second contribution addresses this issue: we present a method called *KrowDD*, that supports data scientists in estimating the relevancy of a feature to their target variable, *before* data for that feature was obtained. *KrowDD* was evaluated on 3 different data sets and significantly outperformed human experts in 2 of them. We therefore find *KrowDD* to yield good results in problems with a balanced target variable, i.e. where both outcomes of a target variable occur equally frequently in the dataset.

Both contributions, the suboptimality of current state-of-the-art as well as *KrowDD* may help raise the awareness for the problem of *data* selection, which could pave the way to improve the situation in practice.

Acknowledgment

We would like to thank Michael Feldman and Michael B. Bühler for their feedback on an early draft of this paper. Additionally, we are grateful to the people participating in our survey (domain experts, data scientists and crowd workers on Amazon Mechanical Turk). This work was supported in part by the Swiss National Science Foundation (SNSF-Project: 200021-143411/1).

References

[1] J. Cheng and M. S. Bernstein, "Flock: Hybrid crowd-machine learning classifiers," pp. 600–611, 2015. [Online]. Available: http://doi.acm.org/10.1145/2675133.2675214

[2] F. Provost and T. Fawcett, "Robust classification for imprecise environments," *Machine learning*, vol. 42, no. 3, pp. 203–231, 2001.

[3] I. Slavkov, B. Zenko, and S. Dzeroski, "Evaluation method for feature rankings and their aggregations for biomarker discovery." *Machine Learning in Systems Biology*, p. 115, 2009.

[4] D. Koller and M. Sahami, "Toward optimal feature selection," *13th International Conference on Machine Learning*, 1996.

[5] B. Efron, "Better bootstrap confidence intervals," *Journal of the American Statistical Association*, vol. 82, no. 397, pp. 171–185, 1987. [Online]. Available: http://www.jstor.org/stable/2289144

[6] J. Bredtmann, C. J. Crede, and S. Otten, "Olympic medals: Does the past predict the future?" *Significance*, vol. 13, no. 3, pp. 22–25, 2016. [Online]. Available: http://dx.doi.org/10.1111/j.1740-9713.2016.00915.x

[7] M. Lichman, "UCI machine learning repository," 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[8] D. Kahneman and A. Tversky, "Prospect theory: An analysis of decision under risk," *Econometrica*, vol. 47, no. 2, pp. 263–291, 1979. [Online]. Available: http://www.jstor.org/stable/1914185

[9] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, vol. 34, no. 1-2, p. 28, 1947. [Online]. Available: + http://dx.doi.org/10.1093/biomet/34.1-2.28

[10] L. V. Hedges, "Distribution theory for glass's estimator of effect size and related estimators," *Journal of Educational Statistics*, vol. 6, no. 2, pp. 107–128, 1981.

[11] A. Tversky and D. Kahneman, "Judgment under uncertainty: Heuristics and biases," pp. 141–162, 1975.

[12] E. L. Krinitzsky, "Earthquake probability in engineeringpart 1: The use and misuse of expert opinion. the third richard h. jahns distinguished lecture in engineering geology," *Engineering geology*, vol. 33, no. 4, pp. 257–288, 1993.

[13] A. Tversky and D. Kahneman, "Availability: A heuristic for judging frequency and probability," *Cognitive psychology*, vol. 5, no. 2, pp. 207–232, 1973.

[14] C. R. M. McKenzie, M. J. Liersch, and I. Yaniv, "Overconfidence in interval estimates: What does expertise buy you?" *Organizational Behavior and Human Decision Processes*, vol. 107, no. 2, pp. 179–191, 2008.

[15] T. Oberlechner and C. L. Osler, "Overconfidence in currency markets," 2008.

[16] R. J. Blendon, C. M. DesRoches, M. Brodie, J. M. Benson, A. B. Rosen, E. Schneider, D. E. Altman, K. Zapert, M. J. Herrmann, and A. E. Steffenson, "Views of practicing physicians and the public on medical errors," *New England Journal of Medicine*, vol. 347, no. 24, pp. 1933–1940, 2002.

[17] E. S. Berner and M. L. Graber, "Overconfidence as a cause of diagnostic error in medicine," *The American journal of medicine*, vol. 121, no. 5, pp. S2–S23, 2008.

[18] P. Tetlock, *Expert political judgment: How good is it? How can we know?* Princeton University Press, 2005.

[19] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[20] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial intelligence*, vol. 97, no. 1-2, pp. 273–324, 1997.

[21] M. Ben-Bassat, "35 use of distance measures, information measures and error bounds in feature evaluation," vol. 2, pp. 773 – 791, 1982. [Online]. Available: //www.sciencedirect.com/science/article/pii/S0169716182020380

[22] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Computing and Communications Review*, vol. 5, no. 1, pp. 3–55, 2001.

[23] J. Surowiecki, *The wisdom of crowds*. Anchor, 2005.

[24] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[25] J. L. Treynor, "Market efficiency and the bean jar experiment," *Financial Analysts Journal*, vol. 43, no. 3, pp. 50–53, 1987.

[26] J. P. Simmons, L. D. Nelson, J. Galak, and S. Frederick, "Intuitive biases in choice versus estimation: Implications for the wisdom of crowds," *Journal of Consumer Research*, vol. 38, no. 1, pp. 1–15, 2011.

[27] J. Y. Zou, K. Chaudhuri, and A. T. Kalai, "Crowdsourcing feature discovery via adaptively chosen comparisons," *arXiv preprint arXiv:1504.00064*, 2015.

[28] B. Nushi, A. Singla, A. Krause, and D. Kossmann, "Learning and feature selection under budget constraints in crowdsourcing," 2016.

[29] D. Kahneman and A. Tversky, "On the psychology of prediction." *Psychological review*, vol. 80, no. 4, p. 237, 1973.

[30] S. M. Herzog and R. Hertwig, "The wisdom of ignorant crowds: Predicting sport outcomes by mere recognition," *Judgment and Decision Making*, vol. 6, no. 1, pp. 58–72, 2011.

[31] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.